Propagation de labels dans un graphe

Projet de Graph Mining

Tanguy Blervaque - Guilhem Prince

Git : https://github.com/GuilhemPrince/graph_mining

Sommaire

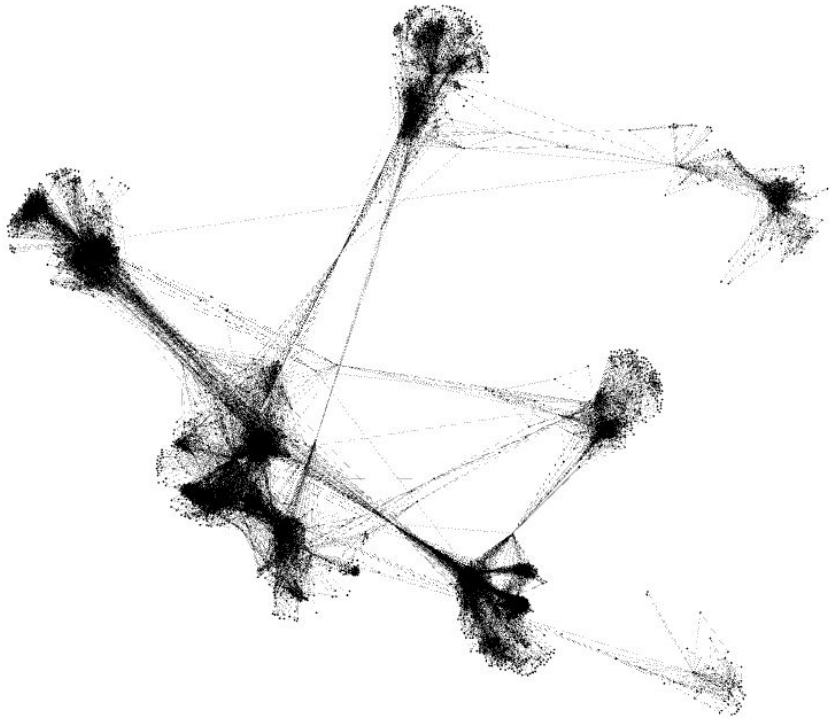
1. Présentation des jeux de données
2. Calcul des centralités
3. Approches de Label Propagation
    a. Approche par apprentissage supervisé
    b. Approche par marche aléatoire
4. Conclusion

1- Présentations des graphes utilisés

‘Facebook social circles’

- Un nœud représente un utilisateur de Facebook, une arête représente l’amitié entre deux utilisateurs.

- Graphe obtenu en combinant les listes d’amis de 10 utilisateurs de Facebook.



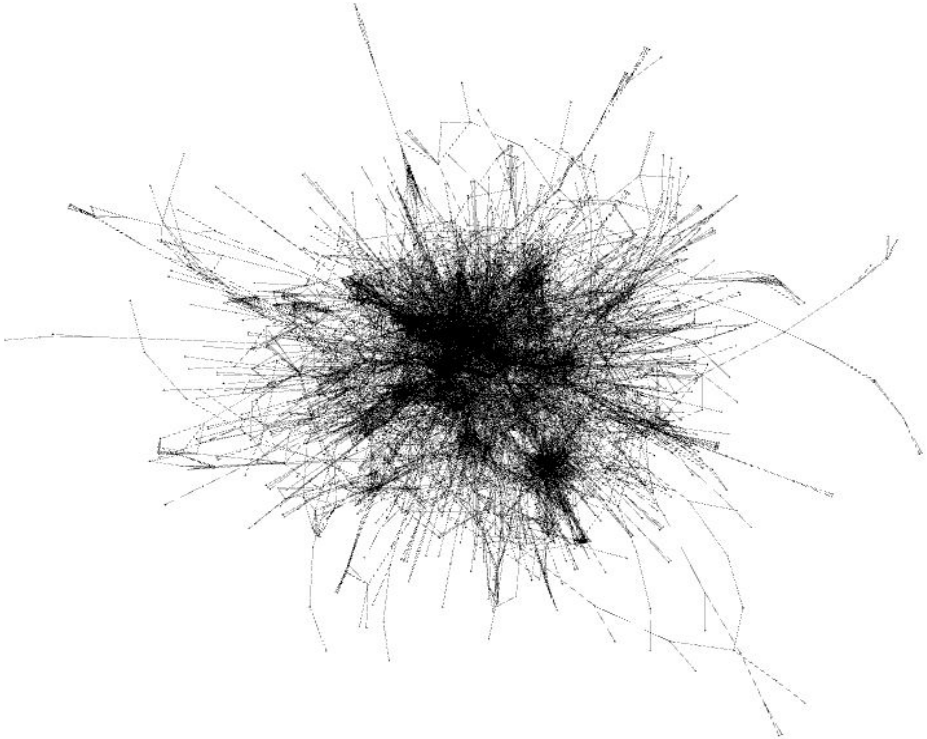
Source : J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012

1- Présentations des graphes utilisés

‘Facebook pages’

- Un nœud représente une page Facebook, une arête représente la présence de likes en commun entre deux pages.

- 8 graphes différents, sur des pages de différents thèmes : Athlètes, artistes, séries TV… Ci-joint, TV shows.



Source : B. Rozemberczki, R. Davies, R. Sarkar and C. Sutton. GEMSEC: Graph Embedding with Self Clustering. 2018.

1- Statistiques des graphes

Source	Dataset	# Nodes	# Edges	Diameter	Density
Fb-circles	—	4,039	88,234	8	0.0108
Fb-pages	TV shows	3,892	17,262	20	0.0023
Fb-pages	Politicians	5,908	41,729	14	0.0024
Fb-pages	Athletes	13,866	86,858	11	0.0009

2- Quelques centralités intéressantes - Fb-circles

Centralité - Rang	Degré	Coefficient de clustering	Betweenness
#1	‘107’ 1045	267 nodes at 1.0	‘107’ 0.48
#2	‘1684’ 792	160 nodes >=90%	‘1684’ 0.34
#3	‘1912’ 755	380 nodes >=80%	‘3437’ 0.24
#4	‘3437’ 547	3232 nodes <80%	‘1912’ 0.23

2- Quelques centralités intéressantes - Fb-pages TV shows

Centralité - Rang	Degré	Coefficient de clustering	Betweenness
#1	‘2008’ 126	333 nodes at 1.0	‘3254’ 0.11
#2	‘3254’ 126	150 nodes >=90%	‘2008’ 0.09
#3	‘3525’ 108	220 nodes >=80%	‘817’ 0.08
#4	‘1177’ 104	3189 nodes <80%	‘2170’ 0.07

3a- Label Propagation par apprentissage supervisé

Méthode :

Approche par classification collective : On applique itérativement la classification sur les voisins sans labels des nœuds avec labels. Les nouveaux nœuds labellisés sont ajoutés à l'ensemble d'entraînement.

sklearn.semi_supervised.LabelPropagation

class `sklearn.semi_supervised.`**`LabelPropagation`**(*kernel='rbf', *, gamma=20, n_neighbors=7, max_iter=1000, tol=0.001, n_jobs=None*) [source]

3b- Label Propagation par Marche Aléatoire

Méthode :

Algorithm 1: $G(V, E)$, labels Y_l

Result: labels $\hat{Y}$

compute $D_{ii} = \sum_j A_{ij}$;

compute $P = D^{-1}A$;

$Y^0 = (Y_l, 0), t = 0$ //Y_u doesn't affect the solution ;

repeat

  $Y^{t+1} \leftarrow PY^t$;

  $Y_l^{t+1} \leftarrow Y_l^t$ //keep the same Y_l;

  $t \leftarrow t + 1$;

until Y^t *converges*;

output Y^t // the most probable label for each node;

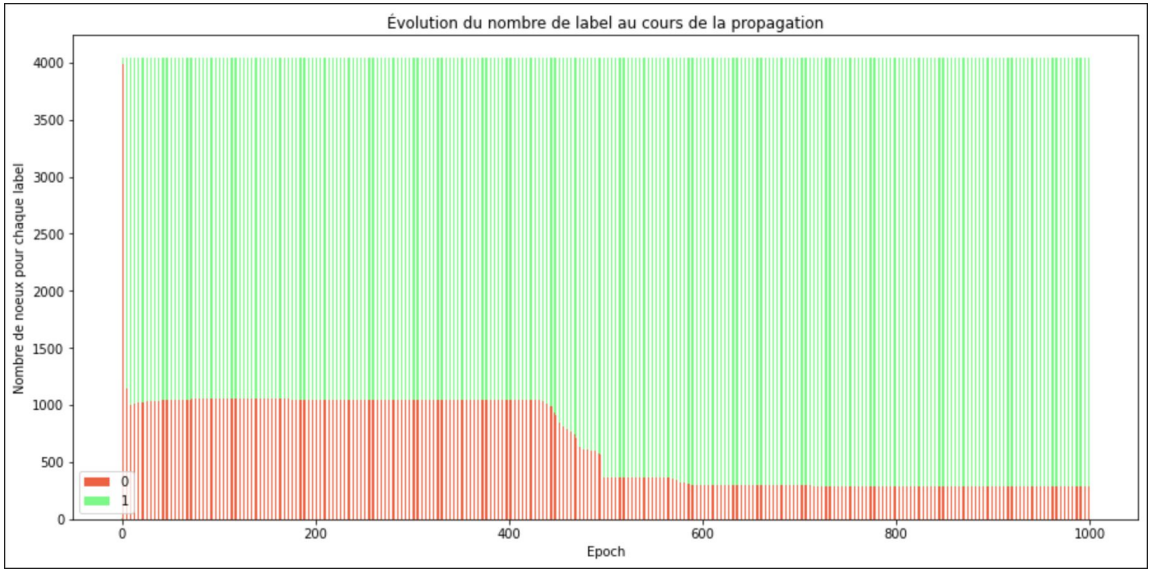
3b- Label Propagation par Marche Aléatoire

Résultats :

On propage deux labels depuis deux couples de 3 noeuds aléatoires ({0: ['19', '87', '347'], 1: ['457', '1', '99']}

Fb-circles

→ Convergence vers un équilibre stable



3b- Label Propagation par Marche Aléatoire

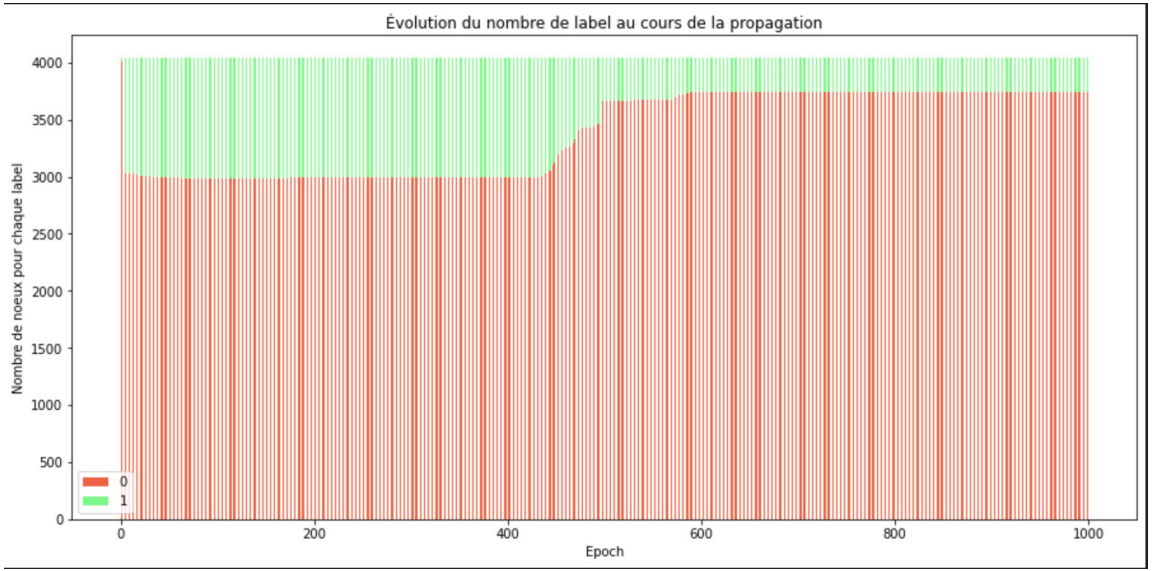
Résultats :

On propage deux labels depuis deux couples de 3 noeuds aléatoires ({0: ['19', '87', '347'], 1: ['457', '1', '99']}

Fb-circles

→ Convergence vers un équilibre stable

→ Inversion des classes lorsqu'on inverse les labels



3b- Label Propagation par Marche Aléatoire

Résultats :

On propage deux labels depuis deux couples de 3 noeuds aléatoires ({0: ['19', '87', '347'], 1: ['457', '1', '99']}

Fb-circles

→ Convergence vers un équilibre stable

→ Inversion des classes lorsqu'on inverse les labels

```
Répartition des labels après label propagation : {0: 291, 1: 3748}
Répartition des clusters après clustering : {0: 203, 1: 3836}
Il y a 494 noeuds dissidents
Il y a 291 noeuds labels 0 qui ne sont pas dans le cluster 0
Il y a 203 noeuds labels 1 qui ne sont pas dans le cluster 1
```

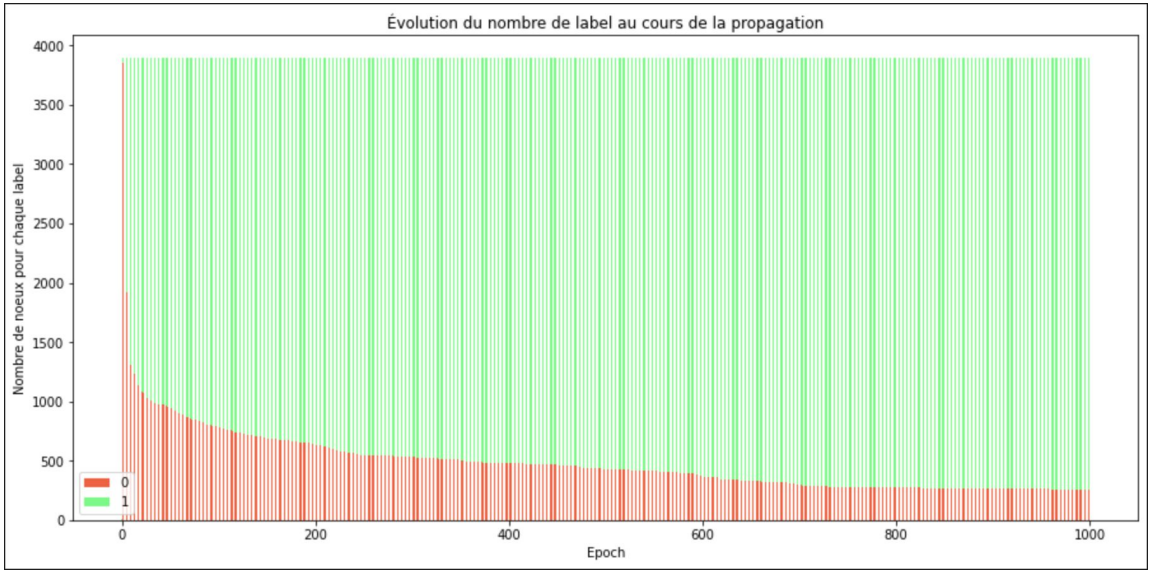
3b- Label Propagation par Marche Aléatoire

Résultats :

On propage deux labels depuis deux couples de 3 noeuds aléatoires ({0: ['19', '87', '347'], 1: ['457', '1', '99']}

Fb-pages tv

→ Convergence vers un équilibre stable



3b- Label Propagation par Marche Aléatoire

Résultats :

On propage deux labels depuis deux couples de 3 noeuds aléatoires ({0: ['19', '87', '347'], 1: ['457', '1', '99']}

Fb-pages tv

→ Convergence vers un équilibre stable

```
Répartition des labels après label propagation : {0: 259, 1: 3633}
Répartition des clusters après clustering : {0: 64, 1: 3828}
Il y a 323 noeuds dissidents
Il y a 259 noeuds labels 0 qui ne sont pas dans le cluster 0
Il y a 64 noeuds labels 1 qui ne sont pas dans le cluster 1
```

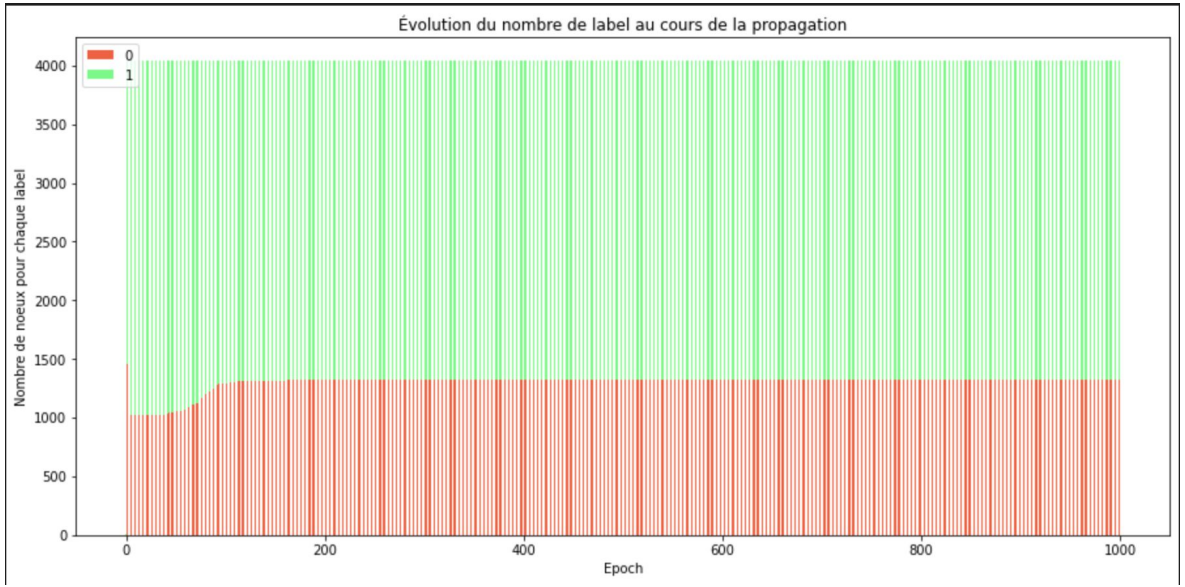
3b- Label Propagation par Marche Aléatoire

Résultats :

On propage deux labels depuis le premier noeud en termes de betweenness centrality vs les 9 suivants

Fb-circles

→ Comme vu lors de l'analyse des centralités, le noeud 107 est très influent



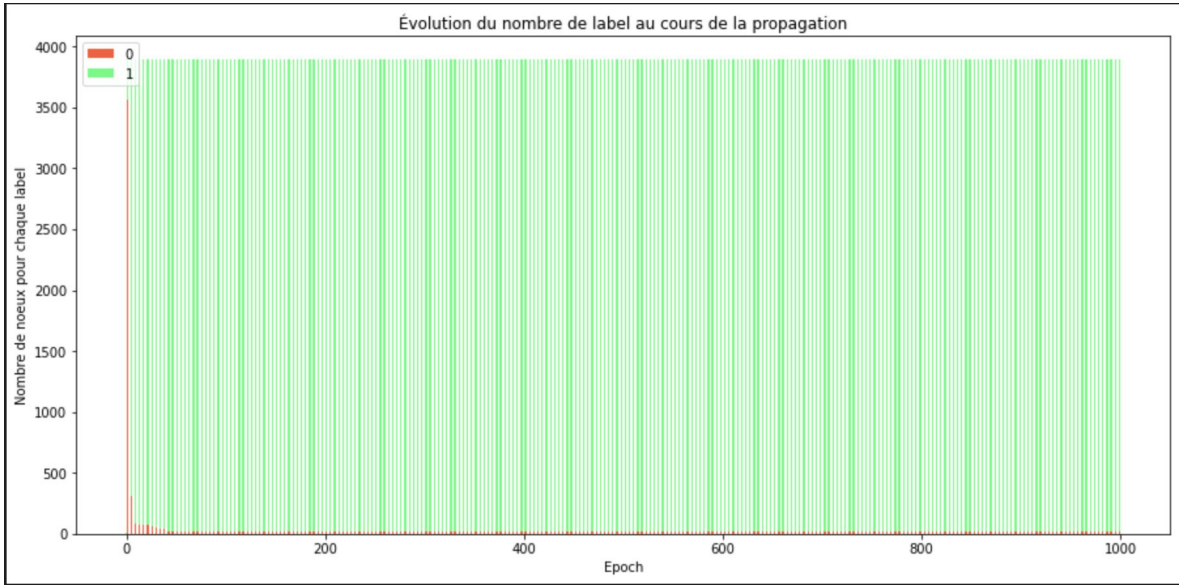
3b- Label Propagation par Marche Aléatoire

Résultats :

On propage deux labels depuis le premier noeud en termes de betweenness centrality vs les 9 suivants

Fb-pages tv

→ À l'inverse, dans ce graphe le noeud 3254 n'est pas si dominant, donc il ne propage pas bien son label



3b- Label Propagation par Marche Aléatoire

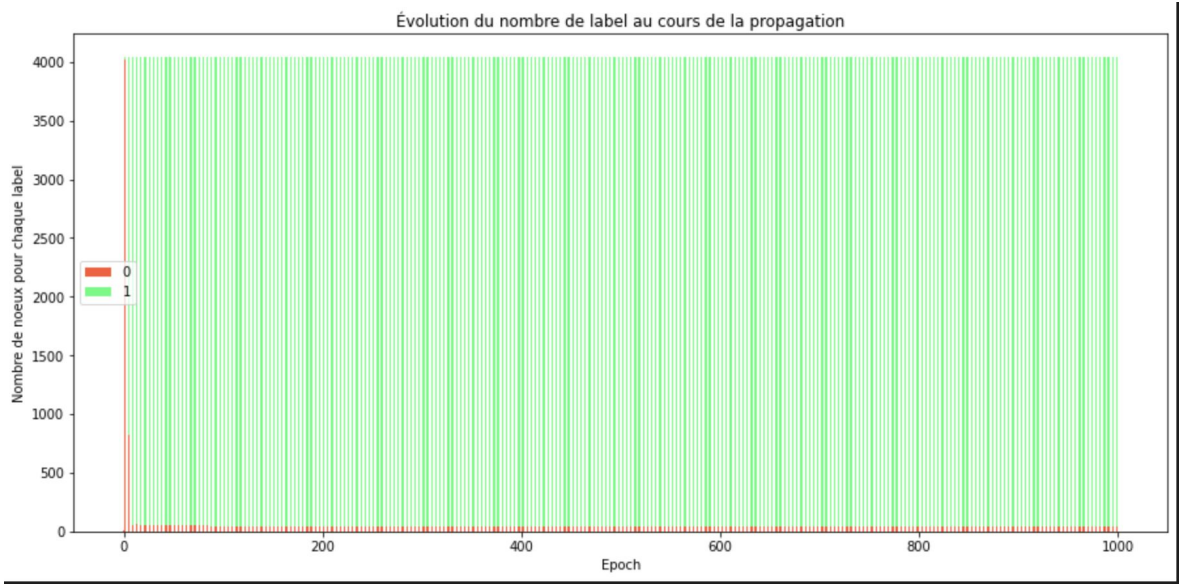
Résultats :

On propage deux labels depuis les 100 premiers noeud en termes de clustering coefficient centrality :
{0: 50 noeuds pris au hasard, 1: 50 autres}

Fb-circles

→ Alors que chaque label est propagé par 50 noeuds initiaux, l'un s'impose largement devant l'autre

→ Due à la structure du graphe



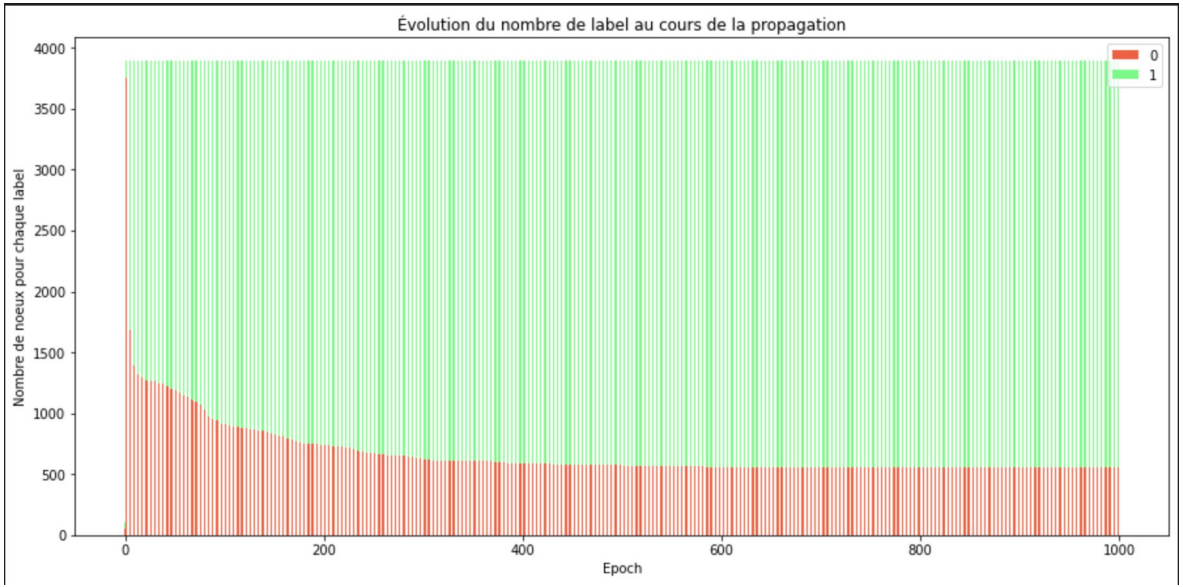
3b- Label Propagation par Marche Aléatoire

Résultats :

On propage deux labels depuis les premiers noeud en termes de clustering coefficient centrality
{0: 50 noeuds pris au hasard, 1: 50 autres}

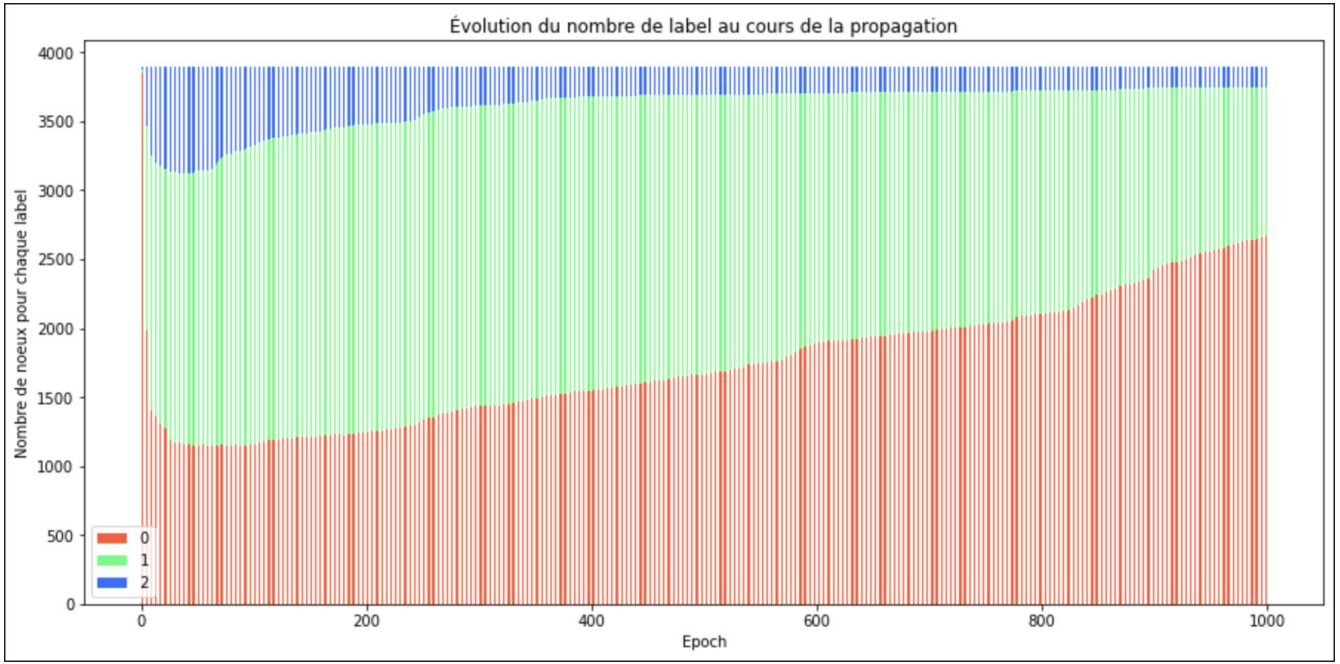
Fb-pages tv

→ Ici le graphe est connecté de manière plus homogène, donc on tend vers un équilibre plus propagé



3b- Label Propagation par Marche Aléatoire

Résultats :



Conclusion

Deux graphes de natures très différentes :

- un graphe hyper clusterisé, réseau d'amis de 10 personnes indépendantes.
- un graphe assez connecté, avec des clusters moins apparents à première vue.

Différentes choses influencent la propagation de labels au sein d'un graphe :

- La structure même du graphe.
- Les noeuds initiaux desquels partent les labels.

Sources : Graphes utilisés

Nous avons utilisé la base de graphes du Stanford Large Network Dataset Collection : https://snap.stanford.edu/data/index.html, en particulier :

- Facebook-circles : https://snap.stanford.edu/data/ego-Facebook.html
- Facebook-pages : https://snap.stanford.edu/data/gemsec-Facebook.html